УДК 004.382.2:004.451.42:.44:004.023

Инструментарий решения масштабируемых задач на распределённых вычислительных системах

Е. Н. Перышкова, А. В. Ефимов, С. Н. Мамойленко[1]

В работе рассмотрена эффективность использования масштабируемых задач на распределённых вычислительных системах (ВС). Авторами предложен способ описания паспорта масштабируемой задачи и создан инструментарий для их обслуживания на основе системы управления ресурсами (СУР) PBS/Torque и планировщика Maui. Проведено экспериментальное исследование эффективности разработанного инструментария при функционировании ВС в мультипрограммном режиме решения наборов масштабируемых задач. Результаты исследования показали, что использование масштабируемых задач по сравнению с задачами с фиксированными параметрами запуска позволяет сократить время решения набора на 25 % и среднее время ожидания задач в очереди СУР на 15 %.

Ключевые слова: распределённые вычислительные системы, управление ресурсами, масштабируемые задачи.

1. Введение

Распределённые вычислительные системы (ВС) относятся к перспективным средствам обработки информации и используются при решении сложных задач науки и техники [1].

В архитектурном плане ВС представляет собой композицию множества элементарных машин (ЭМ) и сети связей между ними. Элементарная машина является основным структурным и функциональным компонентом ВС. Структура ЭМ допускает варьирование от процессорного ядра до конфигураций, включающих универсальные процессоры и специализированные ускорители.

По статистике, 86.8 % ВС из списка top500 [2] являются кластерными. Кластерные ВС функционируют под управлением системного программного обеспечения, включающего в том числе систему управления ресурсами (СУР), например, PBS/Torque [3], Altair PBS Pro, SLURM, HTCondor и др.

Основным назначением ВС является обработка (решение) задач. Под задачей будем понимать требование выполнить параллельную программу на ресурсах ВС. Исследования показывают, что 98 % вычислительных задач обладают возможностью адаптироваться под доступные конфигурации ресурсов ВС перед началом решения [4]. Такие задачи называют масштабируемыми (moldable) [8].

Повысить эффективности эксплуатации ресурсов ВС и снизить время нахождения задач в очереди СУР возможно за счёт разработки инструментария решения масштабируемых задач на распределенных вычислительных системах. Обзор существующих СУР показывает, что в некоторых из них имеется возможность указать лишь диапазон допустимых размеров подсистем ВС для решения масштабируемых задач без учёта пользовательских приоритетов.

Коллективом авторов предложена модификация СУР PBS/Torque и планировщика Maui, реализующая базовые функции по обслуживанию масштабируемых задач, которые описываются набором ресурсных запросов с пользовательским приоритетом.

[1] Работа выполнена при поддержке Совета по грантам Президента Российской Федерации (проект МД-2620-2014.9) и Российского фонда фундаментальных исследований (гранты № 15-07-00048, 13-07-00160).

2. Система управления ресурсами PBS/Torque

В рамках данной работы проведена модификация наиболее распространенного свободно распространяемого программного обеспечения с открытым исходным кодом: PBS/Torque – Terascale Open-Source Resourceand and QUEue Manager.

PBS/Torque – одна из версий системы PBS (Portable Batch System – система пакетной обработки заданий). PBS/Torque управляет загрузкой вычислительных комплексов, состоящих из определенного количества вычислительных узлов, работающих под управлением операционной системы. Функциональная схема СУР PBS/Torque представлена на рисунке 1 [3].

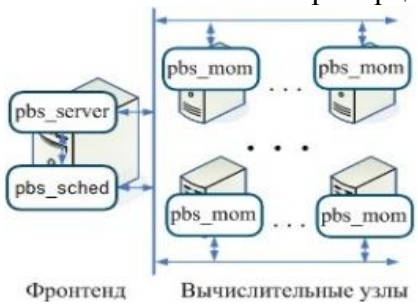

Рис. 1. Функциональная схема СУР PBS/Torque

Каждый узел ВС оснащается локальным компонентом СУР: PBS MOM – Machine Oriented Miniserver, который осуществляет управление вычислительными ресурсами узла. Также предусмотрен централизованный компонент: PBS Server , который поддерживает очередь задач, взаимодействует с планировщиком ресурсов, а также при помощи соответствующего локального компонента координирует работу узлов ВС и осуществляет мониторинг их состояния. В качестве планировщика ресурсов в СУР предусмотрена возможность использовать как встроенный компонент PBS Scheduler , так и внешний модуль, например, Maui Cluster Scheduler.

3. Модульный планировщик MAUI

Maui cluster scheduler — свободно распространяемый планировщик заданий на распределенных вычислительных системах (кластерах) [5]. Maui обладает широким набором настроек и функций, позволяющих выбирать различные политики справедливого планирования, поддерживает резервирование ресурсов и динамическое изменение приоритетов. Широкие возможности по управлению ресурсами ВС и гибкость настройки MAUI позволяет обеспечить эффективное функционирование как простых кластеров, так и суперкомпьютеров при обслуживании пользовательских задач. Модуль планировщика MAUI является альтернативой планировщику pbs_sched, представленному на рис. 1.

4. Паспорт задачи

Задача представляет собой абстрактную сущность, состоящую из набора команд и параметров. Задача представляется пользователем в виде паспорта (скрипта) для оболочки (shell), содержащего атрибуты задачи с описанием требуемых ресурсов и параллельную программу.

```
#PBS -N torque_simple
#PBS -l nodes=1:ppn=4
#PBS -j oe

cd $PBS_O_WORKDIR

echo PBS_NUM_NODES = $PBS_NUM_NODES
echo PBS_NUM_PPN = $PBS_NUM_PPN


/bin/hostname
```

Рис. 2. Паспорт задачи в СУР PBS/Torque

На рис. 2 строки с 1-й по 3-ю содержат директивы СУР PBS/Torque и атрибуты задачи. При постановке задачи в очередь СУР паспорт задачи считывается построчно до тех пор, пока не будет найдена первая строка, которая не является валидной директивой PBS/Torque. Последующая часть паспорта содержит список команд и исполняемые файлы параллельных программ, которые необходимо выполнить.

5. Описание паспорта масштабируемой задачи

Для описания пользователем вариантов подсистем ВС, допустимых для решения масштабируемых задач, предлагается в паспорте вместо атрибута «–l» использовать новый атрибут «–L».

Пример паспорта масштабируемой задачи представлен на рис. 3.

```
#PBS -N moldable_job
#PBS -L nodes=2@ppn=8@weigth=10@walltime=00:25:00,nodes=5@ppn=4@weight=20@walltime=00:20:00
#PBS -j oe

cd $PBS_O_WORKDIR

mpiexec ./test
```

Рис. 3. Паспорт масштабируемой задачи

В параметрах строки после атрибута «–L» указываются через запятую варианты допустимых размеров подсистем ВС. В описании подсистемы допускается указать nodes – количество вычислительных узлов (ВУ), ppn – количество процессорных ядер на каждом узле, weight – приоритет пользователя для подсистемы и walltime – время, на которое необходимо предоставить подсистему. Параметр nodes является обязательным, остальные указываются по желанию пользователя.

6. Инструментарий решения масштабируемых задач

Инструментарий решения масштабируемых задач на распределенных ВС представлен в виде пакета дополнений, содержащего изменения относительно файлов проектов СУР PBS/Torque и планировщика Maui.

При использовании масштабируемых задач в СУР PBS/Torque стандартный планировщик pbs_sched выбирает тот набор параметров, у которого имеется максимальный приоритет. Далее планировщик распределяет ресурсы для задач по алгоритму FCFS (First Come First Served). При его реализации задачи запускаются в том же порядке, в котором они были поставлены в очередь, как только освобождается достаточное количество ресурсов кластера.

При использовании планировщика Maui выбор подсистем ВС для масштабируемых задач реализован в виде перебора вариантов подсистем в соответствии с выбранной политикой Time, Rank, Weight, Worth или None:

Time – запросы ресурсов для задачи рассматриваются в порядке увеличения времени решения (walltime).
Rank – первым выбирается вариант с минимальным количеством процессорных ядер (nodes * ppn).
Weight – предполагает рассмотрение запросов в порядке уменьшения пользовательского приоритета (weight).
Worth – вычисление ценности запросов по формуле 1:

$$\frac{weight_i}{nodes_i * ppn_i * walltime_i}, \quad (1)$$

после чего запросы рассматриваются в порядке уменьшения ценности.
NONE – запросы рассматриваются в порядке их перечисления в паспорте задачи.

7. Экспериментальное исследование

Тестирование программного инструментария проводилось на ресурсах пространственно-распределённой мультикластерной ВС, созданной совместно Лабораторией вычислительных систем ИФП СО РАН и Центром параллельных вычислительных технологий СибГУТИ [6].

Наборы масштабируемых задач генерировались на основе модели рабочей загрузки, предложенной в работе [7]. Количество задач в наборе варьировалось от 50 до 800 с разным процентным содержанием масштабируемых задач (25, 50, 75 и 100 %) совместно с обычными задачами.

Тестовые наборы запускались на ВС в составе 12 ВУ по 4 процессорных ядра на каждом. При этом оценивалась эффективность разработанного инструментария и стандартного программного обеспечения на основе СУР PBS/Torque и планировщика Maui. Стандартные планировщики pbs_sched и maui для масштабируемой задачи использовали первый из возможных вариантов размеров подсистемы ВС.

Для оценки использования программного инструментария определены несколько критериев: общее время решения набора задач, время ожидания задач набора в очереди и загрузка ресурсов ВС.

Результаты влияния процентного содержания масштабируемых задач на время решения набора задач стандартным планировщиком pbs_sched представлены на рис. 4. Оптимальное соотношение количества масштабируемых задач в наборе должно быть больше 50 %. Если уменьшать процентное содержание масштабируемых задач в наборе, то среднее время ожидания задач в очереди и время выполнения задач набора увеличиваются.

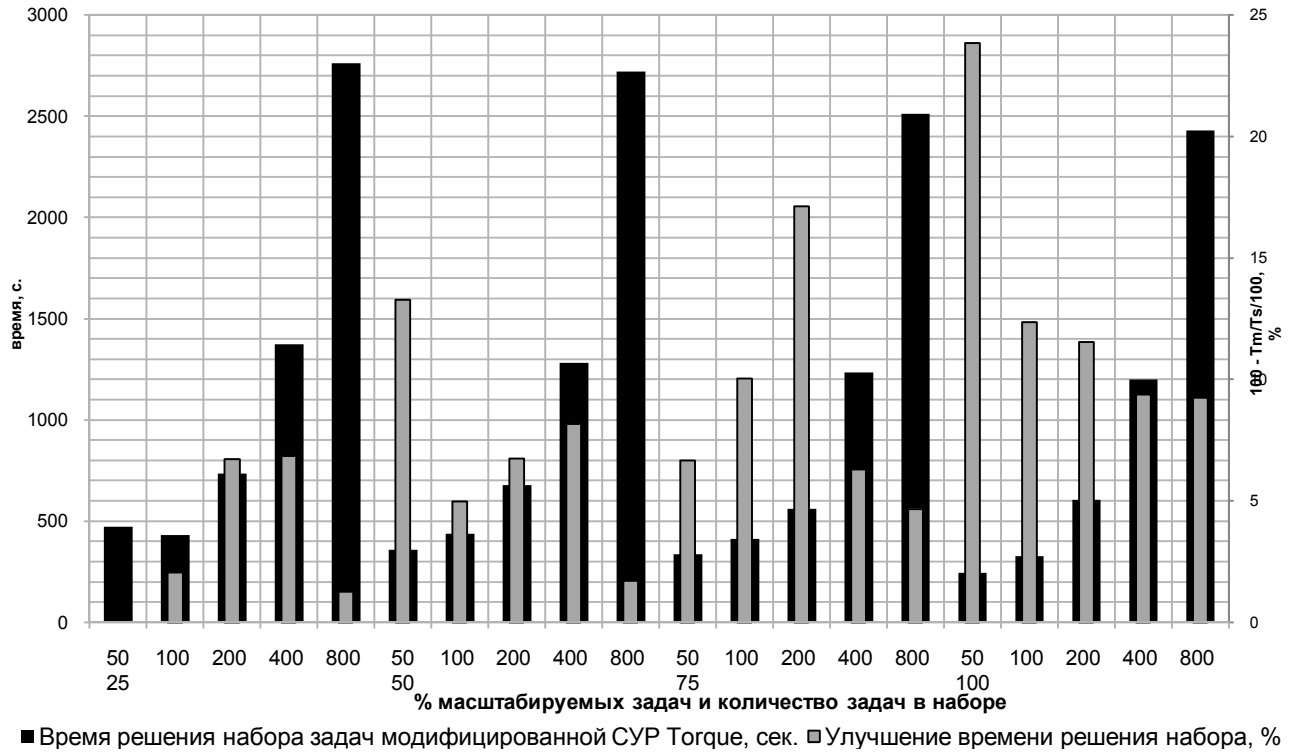

Рис. 4. Исследование влияния процентного содержания масштабируемых задач

При исследовании планировщика Maui внимание уделялось применению различных политик планирования масштабируемых задач и сравнению их по заданным критериям. Результаты экспериментов можно увидеть на рис. 5. При использовании масштабируемых задач можно выделить политики Rank и Worth как наиболее эффективные. Использование данных политик сокращает время решения задач набора в пределах от 15 до 25 % и уменьшает среднее время ожидания задач в очереди на 15 %.

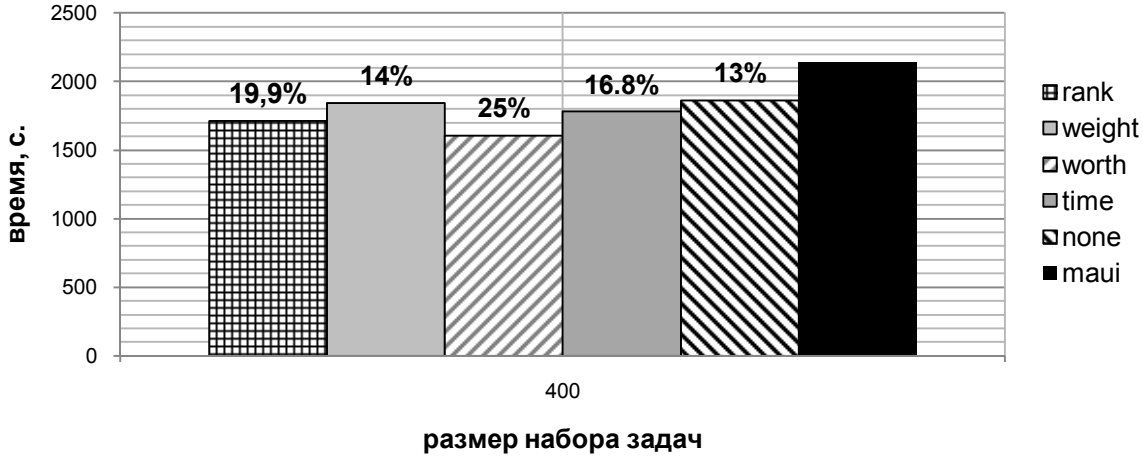

Рис. 5. Исследование политик планирования масштабируемых задач

Для лучших политик планирования проводилось исследование зависимости улучшения времени решения набора задач от количества задач в наборе. Результаты представлены на рис. 6. Применение масштабируемых задач показывает снижение времени решения набора задач и уменьшение времени ожидания задач в очереди на 20 %, начиная с небольших наборов в 50 задач. При увеличении количества задач в наборе эффективность по выбранным критериям растет.

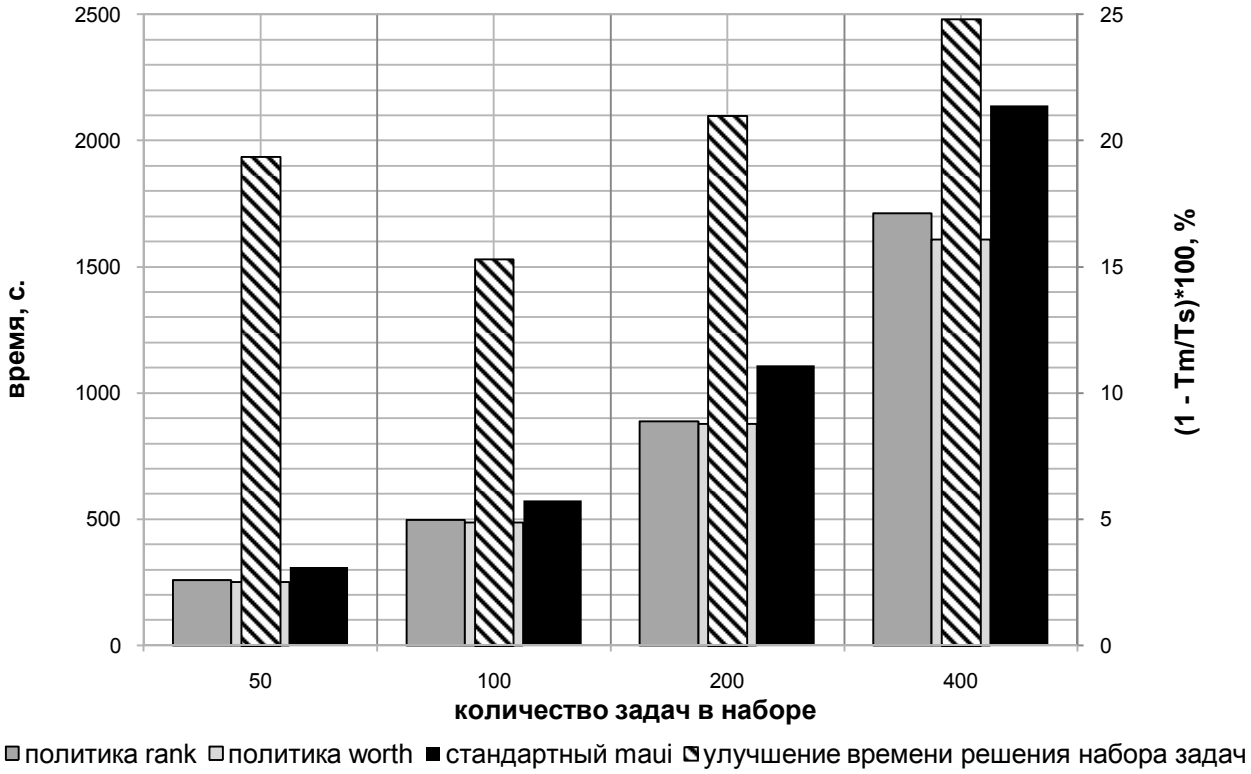

Рис. 6. Исследование зависимости эффективности политик планирования от размера набора масштабируемых задач

Исследование загрузки ресурсов ВС показало, что использование масштабируемых задач не влияет на уровень загрузки ВУ по сравнению с использованием стандартного программного обеспечения. Уровень загрузки ресурсов для различных наборов задач представлен на рис. 7. Процентное значение загрузки при использовании модифицированного планировщика Maui при использовании политики FCFS не превышает 80 % при небольших наборах масштабируемых задач, при наборах задач более 100 загрузка ресурсов ВС не превышает 45 %. Такие данные справедливы и для стандартного планировщика Maui.

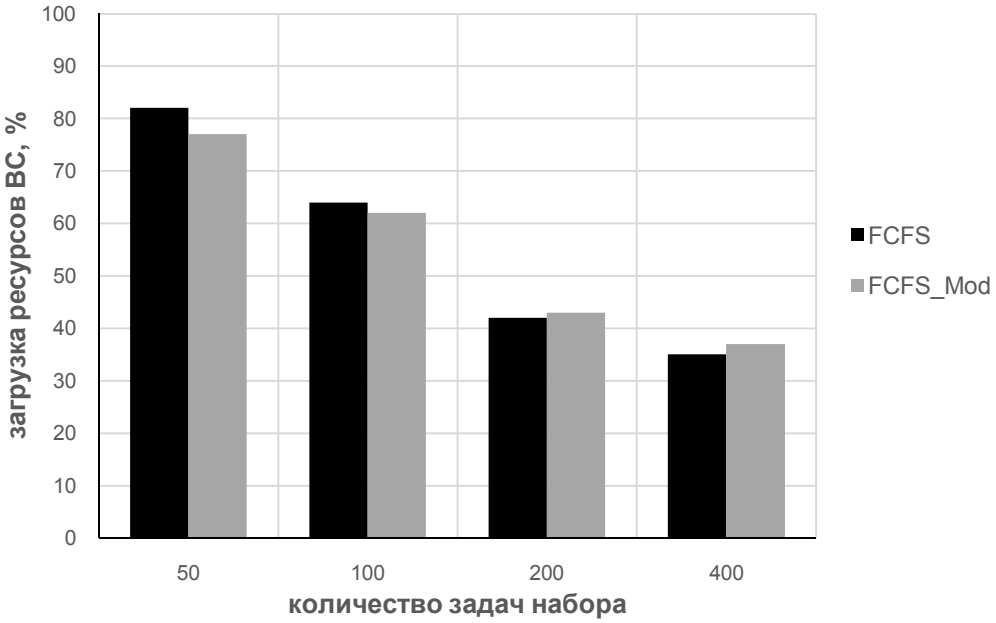

Рис. 7. Исследование влияния масштабируемых задач на загрузку ресурсов ВС

8. Заключение

При использовании стандартного планировщика pbs_shed разработанный инструментарий показал снижение среднего времени решения наборов задач от 10 до 20 % и времени ожидания задач в очереди на 10 % в среднем при использовании 100 % масштабируемых задач в наборе. Использование планировщика Maui дает снижение времени решения наборов масштабируемых задач от 15 до 25 %, а время ожидания задач в очереди снижается на 15 %. Оптимальными политиками планирования являются Rank и Worth.

Литература

1. *Хорошевский В. Г.* Распределённые вычислительные системы с программируемой структурой// Вестник СибГУТИ. 2010. №2 (10). С. 3–41.
2. Top500 super computing sites [Электронный ресурс]. URL: http://www.top500.org/lists/2015/06/ (дата обращения 12.10.2015).
3. Torque Resource Manager [Электронный ресурс]. URL: http://www.adaptivecomputing.com/products/opensource/torque (дата обращения 23.08.2015).
4. *Lifka D.* The ANL/IBM SP scheduling system // Job Scheduling Strategies for Parallel Proc. LNCS. Springer-Verlag, 1995. Vol. 949. P. 295–303.
5. Maui [Электронный ресурс]. URL: http://www.adaptivecomputing.com/products/open-source/maui/ (дата обращения 08.11.2015).
6. Ресурсы центра параллельных вычислительных технологий ФГОБУ ВПО "СибГУТИ" [Электронный ресурс]. Режим доступа: http://cpct.sibsutis.ru/index.php/Main/Resources (дата обращения 15.08.2015)
7. *Cirne W., Berman F.* A model for moldable supercomputer jobs. 15th Intl. Parallel & Distributed Processing Symp. 2001.
8. *Feitelson, D. G.* Toward convergence in job schedulers for parallel supercomputers / D. G. Feitelson, L. Rudolph // Job Scheduling Strategies for Parallel Processing, Lecture Notes in Computer Science. – 1996. – Vol. 1162. – P. 1–26.

Статья поступила в редакцию 30.11.2015

Перышкова Евгения Николаевна
старший преподаватель кафедры вычислительных систем СибГУТИ (630102, Новосибирск, ул. Кирова, 86), тел. (383) 269-82-93,
e-mail: e_perishkova@csc.sibsutis.ru.

Ефимов Александр Владимирович
к.т.н., доцент кафедры вычислительных систем СибГУТИ (630102, Новосибирск, ул. Кирова, 86), тел. (383) 269-82-93, e-mail: efimov@csc.sibsutis.ru.

Мамойленко Сергей Николаевич
д.т.н., доцент, профессор кафедры вычислительных систем СибГУТИ (630102, Новосибирск, ул. Кирова, 86), тел. (383) 269-82-93, e-mail: msn@sibsutis.ru.

Toolkit for moldable jobs management in distributed computing systems

E.N. Peryshkova, A.V. Efimov, S.N. Mamojlenko

This paper considers performance of moldable jobs scheduling in distributed computing systems. We proposed a method for describing moldable jobs scripts and create a toolkit for such job management. The toolkit is based on the resource manager PBS/Torque and the Maui scheduler. We study the queue waiting time and makespan during the servicing sets of moldable jobs. The results of the study show that usage of moldable jobs allows us to reduce the average job queue waiting time by 15% and makespan by 25%.

Keywords: high-performance computing, resource management, moldable jobs.